Can Feature Predictive Power Generalize? Benchmarking Early Predictors of Student Success across Flipped and Online Courses

Mirko Marras
EPFL
mirko.marras@acm.org

Julien Tuân Tu Vignoud
EPFL
julien.vignoud@epfl.ch

Tanja Käser
EPFL
tanja.kaeser@epfl.ch

ABSTRACT

Early predictors of student success are becoming a key tool in flipped and online courses to ensure that no student is left behind along course activities. However, with an increased interest in this area, it has become hard to keep track of what the state of the art in early success prediction is. Moreover, prior work on early success prediction based on clickstreams has mostly focused on implementing features and models for a specific online course (e.g., a MOOC). It remains therefore under-explored how different features and models enable early predictions, based on the domain, structure, and educational setting of a given course. In this paper, we report the results of a systematic analysis of early success predictors for both flipped and online courses. In the first part, we focus on a specific flipped course. Specifically, we investigate eight feature sets, presented at top-level educational venues over the last few years, and a novel feature set proposed in this paper and tailored to this setting. We benchmark the performance of these feature sets using a RF classifier, and we provide and discuss an ensemble feature set optimized for the target flipped course. In the second part, we extend our analysis to courses with different educational settings (i.e., MOOCs), domains, and structure. Our results show that (i) the ensemble of optimal features varies depending on the course setting and structure, and (ii) the predictive performance of the optimal ensemble feature set highly depends on the course activities.

Keywords

Flipped Classroom, MOOC, Success Prediction, Early Warning, Clickstream, At-Risk Students, Learning Analytics.

1. INTRODUCTION

An increasing number of universities are now running blended courses that combine traditional lectures with online instruction, providing educational models tailored to the needs of our society [20]. A popular instructional strategy to enable blended learning is represented by *flipped classrooms*, where students complete *pre-class activities* before attending face-to-face lessons [18]. Recent studies have shown the positive impact and dependency of this strategy on student-centered variables such as self-efficacy and self-regulation [22, 17, 6, 19]. Pre-class activities usually consist in watching videos and completing quizzes part of *Massive Open Online Courses* (MOOCs) used as supplementary material [27]. Each week, students are asked to perform these pre-class activities and to then complete exercises and have discussion in class. Pre-class activities are fundamental for the success of flipped courses [12, 21, 28]. However, students often lack skills, time, and motivation to regulate their pre-class activity; as a consequence, they may experience difficulties in class and end up failing the course [10, 14]. To ensure that no learner is left behind, *Early Success Predictors* (ESPs) are becoming crucial to support instructors in identifying and timely acting upon risk factors of failing a course.

So far, there are few studies on analyzing student success in flipped courses based on pre-class activities. For instance, Jovanovic et al. [9, 8] clustered interaction sequences in pre-class clickstreams to identify learning strategies, showing how strategy-based student profiles differ in course grades. Beatty et al. [2] found that frequency counts of video usage are often correlated with course grades in flipped classrooms. In blended, but not flipped settings, Akpinar et al. [1] showed that student's strategy counts, with strategies modelled as clickstream event n-grams, are indicative of course homework grades. Wan et al. [25, 26] trained gradient boosting classifiers on an extensive set of clickstream-based features to identify at-risk students in a small private online course delivered in hybrid mode. They also analyzed the importance of the features, finding that the time spent in online activities and the stability of time distribution during weeks have the highest importance in that course. To the best of our knowledge, no prior work on flipped courses specifically focused on ESPs.

Conversely, there is a large body of research on success prediction for fully online courses (e.g., MOOCs). A multitude of feature sets have been extracted from clickstreams for this purpose. Recent work proposed video-counting (e.g., number of videos viewed per week, rewinds, fastforwards, pauses, and plays, and the fractional and total amount of time played and paused for videos) and session-based (e.g., number of sessions, mean and standard deviation of the time for all sessions and between sessions) features [4, 13]. These features were fed into different commonly used classifiers

(e.g., Logistic Regression, Naive Bayes, Decision Tree, RF, and Neural Networks) to predict success in weekly assignments or in the entire specific MOOC. In [3], several features that measure intra-course, intra-week and intra-day regularity in video watching were proposed, and their correlation with the course grade was shown. Other researchers leveraged attendance rates, usage rates, and watching ratios [7, 15]. Specifically, they explored how the difference in these indicators affects academic performance, showing that students whose indicators are high are more likely to graduate on schedule. More fine-grained features on video usage (e.g., total video views, mean and standard deviation of the proportion of videos watched, re-watched, and interrupted per week, and the frequency and total number of all video actions and of each type of video action) were proposed in [11]. The authors clustered students according to their watching behavior and found that such a behavior is representative of course performance. Similarly, Mubarak et al. [16] extracted implicit features from video-clickstream data, and investigated the extent to which neural networks fed with those features can predict weekly students' performance. For an extensive discussion on success prediction in MOOCs, we recommend this survey [5].

The above features and classifiers, however, are designed for fully-online learning contexts, such as MOOCs. Despite clear connections, there are essential aspects which distinguish flipped courses from MOOCs. First of all, flipped course data includes relatively few students. A large part of the learning activity happens offline and cannot be tracked, leading to data only on course segments. Flipped courses generally have also an intense instructor guidance and performance on them has direct impact on the academic portfolio. As a motivating example, we consider a flipped course on Linear Algebra later described in this paper and the regularity features proposed for MOOCs in [3]. They quantify students' time regularity by considering their activities over the course (e.g., studying at the same days of the week). Boroujeni's study revealed that the final grade in the MOOC is correlated with two intra-week regularity measures and the periodicity of day hour and week hour ($.46 < c < .7$, $p < 0.001$). Conversely, the same features resulted to have no correlation with the final grade in the above flipped course ($.0 < c < .1$, $p < 0.001$). Therefore, it remains unexplored whether existing features and classifiers for MOOCs generalize to different educational settings (e.g., flipped classrooms), and to what extent the feature importance varies according to the topic, structure, and educational setting of the course.

The contribution of this paper is two-fold: we tackle the problem of ESPs in flipped classroom settings[1], and we provide an extensive analysis and benchmark of classifiers and features for early success prediction across different types of courses, namely MOOCs and flipped courses. A schematic overview of our analysis in this paper is shown in Figure 1.

In a first step, we propose a novel feature set for early success prediction in flipped courses. Our feature set measures students' alignment, anticipation, and strength in quiz and video usage. We benchmark our new feature set using a Random Forest (RF) classifier against eight feature sets presented in previous work on success prediction in online courses. We retrieved these feature sets by systematically scanning the recent papers published at major educational venues (e.g., EDM, AIED, etc.) and reproducing the features based on the relevant papers. Our results on data of 214 students enrolled in a linear algebra flipped course show that the novel feature set outperforms all previously suggested feature sets. We also show that predictive performance can be increased by selecting the optimal features from the ensemble of all feature sets.

In a second step, we extend our analysis to further courses along three dimensions: domain, structure, and educational setting. We compute the early predictive performance again using a RF classifier for three additional courses: a flipped course on functional programming (where pre-class activities include videos only), a MOOC on linear algebra (including video and quiz activities), and a MOOC on functional programming (including video activities only). For each course, we select the optimal features from the ensemble of feature sets (eight feature sets from prior work and one novel feature set from this paper) as input features for the RF classifier. Our results show that the structure of the course significantly influences performance. Predictive performance for courses including quizzes is much higher than for courses including only videos. Furthermore, we also show that while there is some overlap between the optimal features across courses, the importance of the features highly depends on the setting and structure of the course.

2. EARLY PREDICTION FORMULATION

The problem addressed in this paper can be framed into a time series classification task that relies on clickstreams to predict student success in a course. For clarity and reproducibility, we present and formalize the addressed problem.

Course. Early success predictions are provided in the context of a course (e.g., a MOOC or a course run in a flipped classroom setting). In what follows, we hence mathematically define fundamental concepts, such as the course schedule, the learning objects, and their properties. Specifically, we consider a set of students $\mathbb{U}$ who are enrolled in a course c part of the online educational offering $\mathbb{C}$. Each course $c \in \mathbb{C}$ has a pre-defined schedule $\mathbb{S}_c$ consisting of $N = |\mathbb{S}_c|$ online activities, such that $\mathbb{S}_c = \{s_1, \ldots, s_N\}$. We assume that each online activity s_j included in the course schedule is represented by a tuple (o_j, t_j), consisting of learning object $o_j \in \mathbb{O}$ and its corresponding completion deadline for students $t_j \in \mathbb{R}^+$, modelled as a timestamp. Each learning object $o \in \mathbb{O}$ is characterized by descriptive properties denoted with an M-dimensional vector $f_o = (f_1, \ldots, f_M)$ over a set of features $\mathbb{F} = \{\mathbb{F}_1, \ldots, \mathbb{F}_M\}$ that vary according to the type of the learning object (e.g., the duration for a video or the maximum grade for a quiz). Specifically, each feature $\mathbb{F}_j \in \mathbb{F}$ can be envisioned as a set of discrete or continuous values describing an attribute of a learning object o, $f_{o,j} \in \mathbb{F}_j$ for $j = 1, \ldots, M$. Our study in this paper assumes that learning objects can be either videos or quizzes, but the notation can be easily extended to other types (e.g., forum posts or readings). The type of a learning object $o \in \mathbb{O}$ is returned by a function $type : \mathbb{O} \rightarrow \{video, quiz\}$.

[1] https://github.com/d-vet-ml4ed/flipped-classroom

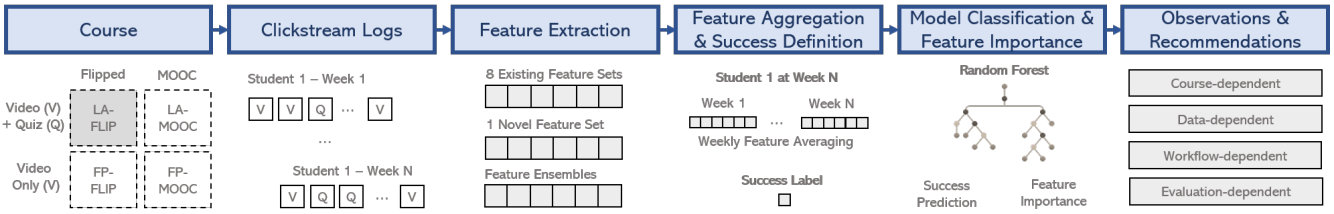


Figure 1: Our Framework. We first analyze a flipped course with videos and quizzes, then investigate differences between courses in flipped and MOOC settings and with videos only and videos plus quizzes. Eight state-of-the-art feature sets, a novel feature set, and feature ensembles are computed for each student and each week of the course. Weekly features are averaged and a success label is attached, according to the course type. Classification is performed using a Random Forest. Observations and recommendations on the predictive power of features are provided for each course setting, highlighting open challenges.

Based on common log data collected by educational platforms, we assume that learning objects of type *video*, denoted as $\mathbb{O}^{video} = \{o \in \mathbb{O} \,|\, \text{type}(o) = video\}$, are described by properties associated to the video duration in seconds as $\mathbb{F}^{video} = (\text{duration} \in \mathbb{R}^+)$. Learning objects of type *quiz*, denoted as $\mathbb{O}^{quiz} = \{o \in \mathbb{O} \,|\, \text{type}(o) = quiz\}$, are characterized by descriptive properties that model the maximum grade students can achieve in that quiz as $\mathbb{F}^{quiz} = (\text{maxgrade} \in \mathbb{R}^+)$. For convenience, we use superscripts to denote a descriptive property of a learning object. For instance, the duration of a video $o \in \mathbb{O}^{video}$ can be referred to as $o^{duration}$. The same notation applies to other descriptive properties.

Interaction. Students enrolled in an online course interact with the learning objects included in the course schedule, generating a time-wise clickstream. We denote a clickstream in a course $c \in \mathbb{C}$ for a student $u \in \mathbb{U}$ as a time series I_u, such that $I_u = \{i_1, ..., i_K\}$, with $K \in \mathbb{N}$ (e.g., a sequence of video plays and pauses, quiz submissions, and so on). We leave these definitions very general on purpose, in particular allowing the length of each time series to differ, since our models are inherently capable of handling this. Likewise, we neither enforce nor expect all time series to be synchronized, i.e. being sampled at the same time, but rather we are fully agnostic to non-synchronized observations. This configuration is common in educational time series. We assume that each interaction i_j is represented as a tuple (t_j, a_j, o_j, d_j), consisting of a timestamp $t_j \in \mathbb{R}^+$, the action $a_j \in \mathbb{A}$ performed by the student (e.g., play or pause), the learning object $o_j \in \mathbb{O}$ involved in the action a_j (e.g., a certain video or quiz), and an L-dimensional descriptive vector $d_j = (d_1, \ldots, d_L)$ over a set of features $\mathbb{D} = \{\mathbb{D}_1, \ldots, \mathbb{D}_L\}$. These descriptive vectors are used to append relevant information to an action a_j performed at time t_j, such as the current video time when the action occurred or the grade received by the student on a quiz. Based on the type of the learning object $o \in \mathbb{O}$, the student can perform different actions $\mathbb{A}$. We assume that video interactions, denoted by $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, \text{type}(o_j) = video\}$, are limited to actions $a_j \in \mathbb{A}^{video} = \{Load, Play, Pause, Stop, SpeedChange, Seek\}$. These actions are derived from those commonly allowed to students in online educational platforms. Conversely, quiz interactions, denoted by $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, \text{type}(o_j) = quiz\}$, include actions $a \in \mathbb{A}^{quiz} = \{\text{Submit}\}$.

In online educational platforms, clickstream interactions include a payload with additional information beyond the timestamp, the action, and the involved learning object. For instance, if a student submits a quiz, the resulting interaction includes also the grade assigned by the system to the student's quiz. Our notation models each dimension $\mathbb{D}_l \in \mathbb{D}$ of a clickstream interaction as a set of discrete or continuous values describing the interaction $i_j \in \mathbb{I}_u$, $d_{j,l} \in \mathbb{D}_l$ for $l = 1, \ldots, L$. Specifically, we assume that interactions involving base video actions $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, a_j \in \{Load, Play, Pause, Stop\}\}$ include descriptive properties associated to the current video time the interaction occurred, i.e. $\mathbb{D}^{Base} = (\text{current-time} \in \mathbb{R}^+)$. Interactions involving a speed change in a video, denoted as $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, a_j \in \{SpeedChange\}\}$, are characterized by descriptive properties associated to both the old and the new speed the video has been and will be watched, i.e. $\mathbb{D}^{SpeedChange} = (\text{oldspeed} \in \mathbb{R}^+, \text{newspeed} \in \mathbb{R}^+)$. Interactions generated by students while seeking the video backward or forward, denoted as $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, a_j \in \{Seek\}\}$, are modelled by descriptive properties related to the previous and current video time the student moved on, i.e. $\mathbb{D}^{Seek} = (\text{oldtime} \in \mathbb{R}^+, \text{newtime} \in \mathbb{R}^+)$. Finally, submit interactions generated in quiz activities, denoted as $\{i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \,|\, a_j \in \{Submit\}\}$, include descriptive properties on the grade assigned to the quiz answer and the progressive number of the attempt made on that quiz, i.e. $\mathbb{D}^{Submit} = (\text{grade} \in \mathbb{R}^+, \text{subnum} \in \mathbb{R}^+)$, with $\text{grade} \in [0, 1]$.

For convenience, we denote as $\mathbb{I}_u^t$ the clickstream including interactions $i_j \in \mathbb{I}_u$, such that $t_j < t \; \forall t_j \in \mathbb{I}_u^t$, namely those occurred before time t. Similarly, since online activities in MOOCs and flipped courses are organized on a weekly basis, t_w identifies the time t where the course week w ends. For instance, the clickstream of user u generated till the end of the second week can be denoted as $\mathbb{I}_u^{t_2}$.

Success Label. Once interactions are modelled, we need to associate a success label according to the final grade the corresponding student has received for that course. We consider a dataset $\mathbb{G}$ to consist of tuples, i.e. $\mathbb{G} = \{(I_{u_j}, y_{u_j})\}$, where I_{u_j} denotes the interactions of student u_j and $y_{u_j} \in \{0, 1\}$ the pass-fail label or the above-below average grade label.

Feature Extraction. Machine-learning models rarely receive raw interaction sequences, as so we abstract such interactions through a feature extraction step. Given the interactions $\mathbb{I}_u^{t_w} \subset \mathbb{I}_u \in \mathbb{I}$, generated by student u till the course week $w \in \mathbb{N}$, we produce fixed-length representations in

$\mathbb{H} \subset \mathbb{R}^{w \times H}$, where $H \in \mathbb{N}$ is the dimensionality of the feature set. Therefore, we assume that H-dimensional vectors are extracted for each week. For instance, if the feature set includes "number of sessions" and "number of clicks", feature vectors of size $H = 2$ are extracted each week. Formally, the extraction process is denoted as $\mathcal{H} : \mathbb{I} \rightarrow \mathbb{H}$, from interactions to features.

Model. Given the dataset $\mathbb{G}$ with interactions - success label pairs, an early success predictor $\mathcal{E}$ aims to predict the success label y_{u_j} associated to the interactions $\mathbb{I}_{u_j}$. Formally, this operation can be abstracted as a function $\widetilde{y}_{u_j} = \mathcal{E}(\mathbb{I}_{u_j}|\theta)$, where $\widetilde{y}_{u_j}$ denotes the predicted label, θ denotes the model parameters, and $\mathcal{E}$ denotes the predictive function that maps interactions $\mathbb{I}_{u_j}$ to the predicted label $\widetilde{y}_{u_j}$ according to θ.

Objective Function. Hence, training an early success predictor $\mathcal{E}$ with interactions - success label pairs till course week w becomes an optimization problem, aimed to find model parameters θ that maximize the expectation on the following objective function (i.e., predicting the correct success label, given the interactions) on a dataset $\mathbb{G}$:

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmax}} \underset{(\mathbb{I}_u, y_u) \in \mathbb{G}}{\mathbb{E}} y_u = \mathcal{E}(\mathcal{H}(\mathbb{I}_u^{t_w}) \mid \theta) \tag{1}$$

In this paper, we focus on *feature extraction*, which formally results in the operationalization of the function $\mathcal{H}$.

3. REPRESENTATIVE FEATURE SETS

To make sure that our work is not only based on individual examples of published research, we systematically scanned the proceedings of conferences and journals for relevant papers in a manual process. In our analysis, we considered papers that appeared in the last years in the top educational technology conferences (e.g., LAK, EC-TEL, AIED, and EDM) and journals (e.g., IEEE TLT, Springer EIT, Journal of Learning Analytics). We considered a paper to be relevant if it (a) proposed a novel feature set for course success analysis, and (b) focused on the context of online courses or courses with online activities. Papers on other tasks, e.g., prediction of affective state or conceptual understanding, or other educational contexts, e.g., interactive simulations or games, were not considered. Moreover, papers with highly overlapping feature sets were filtered, and the paper with the most extensive set was used as representative. Finally, eight papers were included in our study.

In a next step, we reproduced the feature sets described in the above papers. Our approach was to rely as much as possible on the artifacts provided by the authors themselves, i.e., their source code and the descriptions included into the papers. In theory, it should be possible to reproduce published results using only the technical descriptions in the papers. In reality, there are many tiny implementation details with an impact on experiments. Overall, we could reproduce with reasonable assumptions all eight feature sets based on the relevant papers. In what follows, we give a description of each feature set included in our study.

AkpinarEtAl. This feature set consists of consecutive subsequences of n clicks extracted from the session clickstreams of a blended course [1]. In addition to sub-sequences, the authors considered four features related to the number of clicks, the number of session clickstreams, and attendance information. Note that in comparison to the original paper, we extract sequences from a different set a raw events, namely only videos and quizzes (e.g., no events on forums). Hence, in our case the feature set has a size of $|\mathbb{A}^{video} \cup \mathbb{A}^{quiz}|^n$ features per student. Since we expect short patterns to be un-interpretable and particularly long patterns to be rare, we choose $n = 3$ for our analyses.

BoroujeniEtAl. This feature set was originally used to measure to what extent MOOC students are regular in their study patterns [3]. Specifically, it is considered whether students study on certain hours of the day, day(s) of the week or similar weekdays. Other features monitor whether students have the same distribution of study time among weekdays over weeks, particular amount of study time on each weekday, and finally to what extent a student follows the schedule of the course. This set includes 9 features per student. Other papers proposed similar regularity features [23, 24, 8, 9, 2]. We limit our analysis to the feature set listed in [3], as in our first experiments it exhibited the best predictive power (among papers focusing on regularity features).

ChenCui. The feature set presented in this paper [4] includes click countings from a mandatory undergraduate course run through Moodle. Features include the number of total clicks and of clicks on campus, the ratio of on-campus to off-campus clicks, the number of online sessions (with average and standard deviation), standard deviation of time between online sessions, number of clicks during weekdays or weekends, ratio of weekend to weekday clicks, and the number of clicks for each type of module (e.g., assignment, forum, and quiz). To accomodate the scenario presented in Section 2, our study does not cover the features not easily generalizable to different types of online courses: the number of clicks on campus, the ratio of on-campus to off-campus clicks, the number of clicks for modules file, forum, report system. We therefore obtain a feature set of size 13 for each student.

LalleConati. This paper [11] focuses again on MOOCs. The presented feature set is composed by video interaction features at two levels of granularity. Features on video views include the total number of videos views (both watches and rewatches), in addition to the average and standard deviation of the proportion of videos watched, re-watched, and interrupted per week. On the other hand, features on actions performed within the videos include the frequency and total number of all performed video actions, frequency of video actions for each type of video action, and the average and standard deviation duration of video pauses, seek lengths, and so on. This feature set has a size of 22 per student.

LemayDoleck. The next paper [13] is also focused on MOOCs. Presented features include the number of videos watched per week, the average time fraction paused, played or spent watching, the average and standard deviation of the playback rate, and the total number of rewinds, pauses, and fast-forwards. Note that this feature set includes only video-related measures, resulting in vectors of size 10 per student.

MbouzaoEtAl. In this MOOC paper [15], the authors introduce three novel features, namely attendance rate, uti-

lization rate, and watching ratio. The attendance rate of a student on a given week is the number of videos the student played over to the total number of videos in that week. The utilization rate is the proportion of video play time activity of the student over the sum of video lengths for all videos on that week. Finally, the watching ratio is defined as the product between the two former features. This 3-sized feature set has been tested in MOOCs, extending an already-existing feature set [7].

MubarakEtAl. This paper [16] is primarily focused on implicit features about video-usage behavior in MOOCs. Composed by 13 features, this set covers fine-grained characteristics, such as the percentage of the video the learner watched not counting repeated segments, the amount of real time the learner spent watching the video (i.e. when playing or pausing) compared to the video duration, and the sum of times a learner viewed a video in its entirety.

WanEtAl. This set was designed for a small private online course [26]. Features measure the online learning time, the strength of the learner's engagement in forums and weekly assignments, the extent to which students attempt to do the homework soon, as examples. Table 1 and 2 in the cited paper provide further details. Given that we do not cover forum interactions, our study does not consider forum features. Finally, this set includes 14 features per student.

4. EARLY PREDICTORS IN FLIPPED COURSE SETTINGS

In this section, we first present a novel feature set for flipped courses, based on alignment, anticipation, and strength in content usage. We then describe the experimental setup and results aimed to assess to what extent the feature sets (including ours) are predictive of student success.

4.1 Our Feature Set

The feature sets presented so far mainly tackle video-related features and/or consider only low-level features, with only a few of them including features related to quizzes or assignments. Considering that predicting the success of a student based on clickstream data only is a challenging task per sé, we believe that limiting features to those extracted from videos may result in inferior predictive performance. We therefore suggest a number of additional features assessing students' knowledge and alignment with the course schedule.

Competency Strength is defined as the average of the inverse number of submissions for a quiz, weighted by the highest grade achieved by the student on that quiz. Given the inverse term, the value of this feature decreases when the student attempts the quiz multiple times and if the grade achieved by the student on the last attempt is not the highest-possible one. Hence, good-performing students may use few attempts and reach the maximum quiz grade fast (value close to 1). Students struggling with the material may attempt the quiz many times and not reach the maximum grade (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{1}{|Q_u|}\sum_{q\in Q_u}\frac{1}{Q_u^q}\,max(G_u^q) \tag{2}$$

where:

- $Q_u = \{o_j | i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \cap \text{type}(o_j) = quiz \cap t_j \leq t_w\}$ are the quizzes taken by student u till week w.
- $Q_u^q = |\{i_j | i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \cap o_j = q \cap t_j \leq t_w\}|$ is the number of attempts a student had on quiz q.
- $G_u^q = \{d_j^{\text{grade}} | i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \cap o_j = q \cap t_j \leq t_w\}$ is the set of grades a student got on quiz q.

Competency Alignment is defined as the number of quizzes the student received the maximum grade until week w, divided by the total number of quizzes scheduled for the period of consideration. Good-performing students may receive the maximum grade in all quizzes for the period of consideration (value close to 1); low-performing students may be behind the schedule and pass fewer quizzes than those proposed (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{|Q_u^{\text{pass}} \cap S^{leq(t_w)}|}{|S^{leq(t_w)}|} \tag{3}$$

where:

- $Q_u^{\text{pass}} = \{o_j | i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \cap \text{type}(o_j) = quiz \cap d_j^{\text{grade}} = o_j^{\text{maxgrade}}\}$ is the set of quizzes the student u received the maximum grade until week w.
- $S^{leq(t_w)} = \{o_j \in \mathbb{O} | (o_j, t_j) \in \mathbb{S}_c \cap type(o_j) = \text{quiz} \cap t_j \leq t_w\}$ is the set of quizzes to complete by week w.

Competency Anticipation is defined as the number of quizzes attempted by the student among those in subsequent weeks of the current week of study. This feature can be seen as a proxy of the learning propensity of a student. For instance, if a quiz is scheduled to be solved in subsequent weeks, we expect that good-performing students try them earlier, anticipating the deadline stated in the platform (value close to 1). Low-performing students may delay the consumption of quizzes across weeks or even towards the end of the course (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{|Q_u \cap S^{gt(t_w)}|}{|S^{gt(t_w)}|} \tag{4}$$

where Q_u is the set of quizzes taken by student u until week w as defined in Eq. 2, and:

- $S^{gt(t_w)} = \{o_j \in \mathbb{O} | (o_j, t_j) \in \mathbb{S}_c \cap type(o_j) = \text{quiz} \cap t_j > t_w\}$ is the set of quizzes to complete after week w.

Content Alignment is defined as the number of videos watched by the student until week w, divided by the total number of videos scheduled for the period of consideration. Good-performing students are expected to complete all videos for the period of consideration (value close to 1), while low-performing students may complete less videos than those proposed (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{|V_u \cap S^{leq(t_w)}|}{|S^{leq(t_w)}|} \tag{5}$$

where:

- $V_u = \{o_j | i_j = (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \cap \text{type}(o_j) = video\}$ is the set of videos watched by student u until week w.

- $S^{leq(t_w)} = \{o_j \in \mathbb{O} | (o_j, t_j) \in \mathbb{S}_c \cap type(o_j) = \text{video} \cap t_j \leq t_w\}$ is the set of videos to watch by week w.

Content Anticipation is defined as the number of videos completed by the student among those in subsequent weeks of the current week of study. For instance, if a video is due the next week, we expect that good-performing students might watch them earlier, anticipating the deadline stated in the platform (value close to 1). On the other hand, low-performing students may tend to delay the completion of videos (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{|V_u \cap S^{gt(t_w)}|}{|S^{gt(t_w)}|} \tag{6}$$

where V_u is the set of videos watched by student u until week w as defined in Eq. 5, and:

- $S^{gt(t_w)} = \{o_j \in \mathbb{O} | (o_j, t_j) \in \mathbb{S}_c \cap type(o_j) = \text{video} \cap t_j > t_w\}$ is the set of videos to watch after week w.

Student Shape is defined as the student's tendency of receiving the maximum grade in a quiz at the first attempt in a row. Good-performing students are expected to consecutively receive the maximum grade in quizzes at the first attempt (value close to 1); students experiencing difficulties may require multiple attempts on each quiz, before getting the maximum grade (value close to 0). Given a student u and the week w of the course, this feature is computed as:

$$\frac{1}{\sum_{(p_i,l_i)\in\mathbb{P}} p_i} \sum_{(p_i,l_i)\in\mathbb{P}} \frac{p_i \cdot l_i}{|\{p_i | (p_i, l_i) \in \mathbb{P} \ \cap \ l_i = 1\}| + \epsilon} \tag{7}$$

where $\mathbb{P} = \{(p_0, l_0), \ldots, (p_n, l_n)\}$ represents a series counting how many quizzes the student consecutively receives the maximum grade ($l_i = 1$) or failed ($l_i = 0$) at the first attempt in a row. For instance, if a student gets the maximum grade for the first five quizzes at the first attempt in a row, then is wrong in two quizzes at the first attempt, and then receives the maximum grade for ten quizzes at the first attempt in a row, $\mathbb{P}$ would be equal to $\{(5, 1), (2, 0), (10, 1)\}$.

Student Speed is defined as the average time passed between two consecutive attempts for the same quiz, among those taken by the student. This feature captures intrinsic behavior of students who take the quiz, spending less time or more time to attempt it, on average. Given a student u and the week w of the course, this feature is computed as:

$$\frac{1}{|Q_u|} \sum_{q \in Q_u} \sum_{i=1}^{|t_q|} \frac{|t_q^i - t_q^{i-1}|}{|t_q|} \tag{8}$$

where Q_u is the set of quizzes taken by student u until week w as defined in Eq. 2, and:

- $t_q = [t_j | (t_j, a_j, o_j, d_j) \in \mathbb{I}_u \ \cap \ o_j = q \ \cap \ t_j > t_{j-1})]$ are timings between trials for u on q, chronologically.

In the rest of the paper, we will refer to our set by *Ours*.

4.2 Experimental Evaluation

In this section, we benchmark our new feature set against the eight feature sets presented in prior work (see Section 3), on early success prediction in flipped courses. For convenience, we will use author-based labels to identify feature sets throughout the paper, but we will be more interested in contrasting the impact of features in those papers based on what they implicitly measure (not based on the authors).

4.2.1 Experimental Setup

Protocol. For each dataset, we applied a train-test evaluation, i.e. parameters were fit on the training data set and the performance of the models was evaluated on the test data set. We performed all experiments using Random Forest (RF) classifiers, known to achieve a good trade-off between prediction accuracy and interpretability. Performance of all models was computed using a **nested** student-stratified (i.e. dividing the folds by students) 10-fold cross validation. The same folds were used for all experiments, across feature sets. We optimized the hyper-parameters of RFs via Grid Search in Scikit-Learn. Specifically, we tuned the following hyper-parameters: number of estimators $(25, 50, 100, 200, 300, 500)$, the maximum number of features $(sqrt, None, log2)$, and the splitting criterion $(gini, entropy)$. More extensive grids were run, but they did not show any substantial improvement. To be precise, we determined the set of optimal hyper-parameters as follows: within each iteration, we ran an inner student-stratified 10-fold cross-validation on the training set in that iteration, and selected the combination of hyper-parameter values yielding the highest accuracy on the inner cross-validation. Note that we trained RFs by weeks: the RF for week w of a given course was trained on data collected up to week w. To obtain the input features for RF for week w, we computed the weekly features for the selected feature set and averaged them.

Data Set: LA-Flip. We consider a Linear Algebra course for undergraduate students taught in a flipped format for 10 weeks at EPFL. Typical pre-class work included a list of video lectures and online quizzes from a Linear Algebra MOOC. The final exam grade, lying between 0 and 6, with 4 as passing threshold, is considered as a measure for students' performance. The repeating students were filtered out, given that their repeated exposure to the material might add a bias to our findings. The final dataset consists of clickstream data from 214 students, with 41% of them failing the course. The study was approved by the university's ethics committee (HREC No. 058-2020/10.09.2020).

4.2.2 Observations

We evaluated the predictive accuracy of RF classifiers trained on the different feature sets extracted from LA-Flip under a binary classification that aims to identify passing and failing students early, as described in Section 2. We further also trained RF classifiers only on the most important features selected from all features (denoted as *EnsembleAll*) and from all features except ours (denoted as *EnsembleButOurs*). Figure 2 reports the balanced accuracy, the area under the ROC curve (AUC), and the individual percentage of passing and failing students correctly identified (recall) for each feature set over all weeks and folds.

The lowest-performing feature sets appear those monitoring students' regularity (orange) and attendance and utilization rates (blue). Hence, a first conclusion we can draw is:

Highlight #1. *Regularity and attendance/utilization features, powerful in MOOCs, do not allow to distinguish passing from failing students in the considered flipped course.*

The feature sets mostly related to video-clicking behavior, such as those from Lemay & Doleck, do not lead to substan-

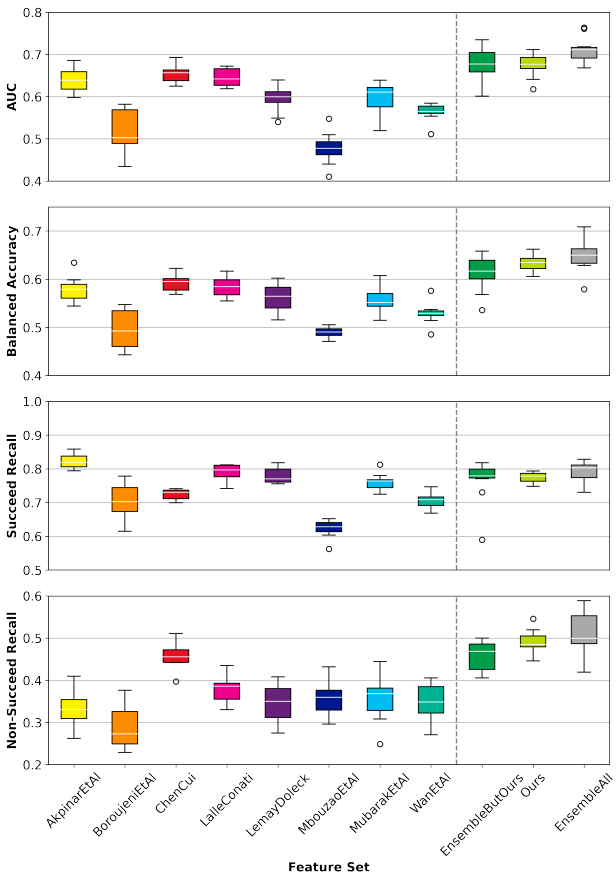


Figure 2: [LA-Flip]. Effectiveness of a RF classifier trained on separate feature sets and on ensembles. Our feature set is essential to increase the effectiveness of the classifier, especially in terms of Non-Succeed (failing students) Recall.

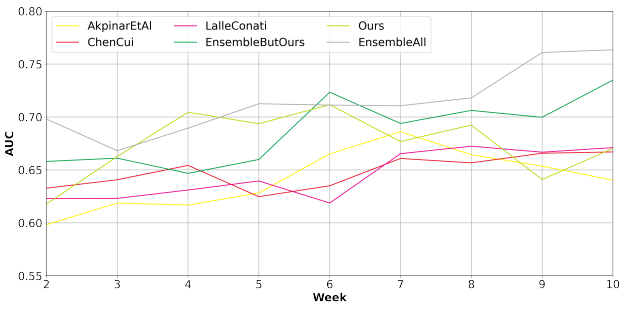


Figure 3: [LA-Flip]. AUC for the best six feature sets. The ensemble of all features (grey) leads to an increase in effectiveness, with respect to considering feature sets separately.

tial differences from each other and all achieved a balanced accuracy between 55% and 59% (similarly for AUC). This finding might reveal an intrinsic limit for video features in predicting student success from pre-class activities. Our results also raise the question on how and why a certain type of video features should be preferred compared to others.

Highlight #2. *In this flipped course, there are minimal differences in performance among video-usage features; an intrinsic predictive limit for video-usage features exists.*

This motivates investigation on the impact of features targeting quiz usage. In this direction, the features proposed by Wan et al. cover a range of raw counting and timing measures that target quizzes. Figure 2 shows that this feature set is even worse that just using video features. Conversely, by measuring more complex patterns in quiz consumption, our feature set led to a balanced accuracy of 67% (similarly for AUC). To identify the aspect our features make the difference at, we considered the percentage of passing and failing students correctly classified, as shown in the two bottom plots in Figure 2. While there are no substantial differences among our feature set and the other ones in identifying passing students (Succeed Recall), a clear improvement is obtained in the detection of failing students (Non-Succeed Recall), fundamental to ensure fewer students are left behind. The impact of our features can be also appreciated across weeks in Figure 3. Our features allowed the ensemble to be effective in the first weeks, while both ours and other features jointly led to an improvement in the second part of the course. Given our results and the characteristics of our features, we can observe that:

Highlight #3. *Extracting fine-grained features that model alignment, anticipation and strength of video/quiz usage results in higher predictive power on failing students.*

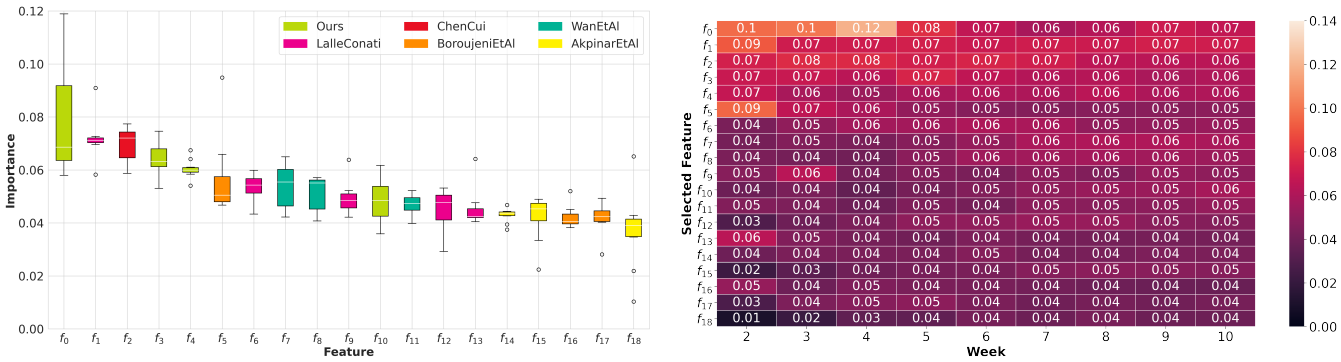
Though considering the feature sets separately allowed us to perform a fine-grained assessment and have an estimation of their predictive power, it remains unclear how the effectiveness of early predictors can be improved by training models with an ensemble of all features and to what extent the importance of the considered features varies. Hence, on the right side of the plots in Figure 2, we present the results achieved by a RF classifier only with the most important features selected from all features and from all features except ours. It can be observed that the optimal ensemble of features without ours results in lower performance, compared to the optimal ensemble that uses also our features. The optimal ensemble of all feature has an AUC score consistently higher than 0.70. To inspect what drives success prediction, we computed the feature importance over weeks and folds, and reported in Figure 4 the importance of features (short description in Table 1) selected by RF. Looking at importance scores in Figure 4(a), we observe that:

Highlight #4. *The extent to which students anticipate content consumption, the tendency of learning during weekends, the proportion of watched videos, and the strength of their performance in quizzes, had the highest importance.*

Figure 4(b) shows that the difference in importance across features is more evident in the first weeks. This finding emphasizes the fact that selecting appropriate features is more crucial when interested in very early predictions.

5. EARLY PREDICTORS OVER COURSES

Our exploratory analysis revealed interesting patterns on the predictive power and importance of a range of features. However, it remains under-explored the extent to which the patterns identified in that flipped course hold also in courses with other structures and educational settings. To this end,



(a) Average feature importance across weeks.

(b) Feature importance over weeks.

Figure 4: [LA-Flip]. Importance of the best nineteen features selected by a RF classifier from the ensemble of all feature sets. Four features of our set have been selected as important. Table 1 lists the Feature IDs and the short description of each feature.

Table 1: [LA-Flip]. Description of the most important nineteen features selected by a RF classifier from the ensemble of all feature sets, showed in decreasing order of importance. Four features of our set have been selected among the top eleven.

ID	Set	Name	Short Description
f_0	`Ours`	CompetencyAnticipation	The extent to which the student approaches soon a quiz provided in subsequent weeks.
f_1	LalleConati	WeeklyPropWatched-Avg	The proportion of videos the student watched, counting repeating segments.
f_2	ChenCui	RatioClicksWeekendDay	The ratio between clicks happened during weekend and weekdays.
f_3	`Ours`	ContentAnticipation	The extent to which the student approaches soon a video provided in subsequent weeks.
f_4	`Ours`	CompetencyStrength	The extent to which a student passes a quiz getting the maximum grade with a low number of trials.
f_5	BoroujeniEtAl	RegWeeklySim-M2	The extent to which the student has a similar distribution of workload among weekdays across weeks.
f_6	LalleConati	WeeklyPropInter-Std	The standard deviation of the time the student spent while interrupting a video, across videos.
f_7	WanEtAl	NumSubmissionCor	The average number of quizzes attempted and correct.
f_8	WanEtAl	NumSubmissions-Avg	The number of submissions required to pass a quiz, on average.
f_9	LalleConati	WeeklyPropInter-Avg	The average time the student spent while interrupting a video, across videos.
f_{10}	`Ours`	StudentShape	The extent to which the student receives the maximum grade in quizzes at the first attempt in a row.
f_{11}	WanEtAl	NumSubmissionPerCorrect	Percentage of the correct quiz submissions with respect to the total submissions.
f_{12}	LalleConati	WeeklyPropReplayed-Avg	The proportion of videos the student re-watched, not counting repeating segments.
f_{13}	LalleConati	FrequencyEvent-VideoPlay	The frequency of the video play action in the students' online sessions.
f_{14}	AkpinarEtAl	QCheck-QCheck-VLoad	The amount of times the student checks twice a given quiz and then go to load a video.
f_{15}	AkpinarEtAl	VPlay-VPause-VLoad	The amount of times the student plays a video, pause and then load the next one.
f_{16}	BoroujeniEtAl	RegPeriodicity-M3	The extent to which the daily study pattern is repeating over weeks (e.g., same days of the week).
f_{17}	BoroujeniEtAl	RegWeeklySim-M1	The extent to which the student works on the same weekdays.
f_{18}	AkpinarEtAl	VStop-PCheck-VLoad	The amount of times the student stops a video, attempts a quiz and then load the next video.

we extended our analysis to a flipped course in a different domain (Functional Programming, only video data in pre-class activities), a MOOC in the same domain (Linear Algebra, both videos and quizzes), and a MOOC from a different domain (Functional Programming, only video interactions).

5.1 Experimental Setup

Protocol. In this experiment, we followed the steps described in Section 4.2.1, with few exceptions. Specifically, for each data set, we considered only classifiers trained with the optimal ensemble of all features proposed in prior work plus the ones proposed in this paper. To obtain the input features for the RF classifier on week w, we computed the weekly features for all feature sets; then averaged features of the same week, and finally averaged across weeks till week w. For each course, we computed the most important features from the ensemble (eight existing sets and ours) based on the average importance of the features across folds and weeks. The study was approved by the university's ethics committee (HREC No. 096-2020/09.04.2020).

Data Set: FP-Flip. We consider one stream of a Functional Programming course taught to EPFL Master's students in a flipped manner for 10 weeks. The preparatory work included a list of videos from a Functional Programming MOOC. Repeating students were filtered out. Being a Master's course with a failing percentage of only 5%, we considered whether a student's final course grade (lying between 0 and 6) was above the average grade over all students as a success label. The dataset consists of clickstreams from 218 students, with 38% of them being below average.

Data Set: LA-MOOC. The content used in pre-class activities within LA-Flip was also provided by EPFL instructors on an external MOOC platform in form of three separate MOOCs, with the first MOOC being equivalent to the first 4 weeks of the flipped course, the second MOOC equivalent to week 5 to week 8, and the third MOOC equivalent to the last 3 weeks. Given that the first 4 weeks of LA-Flip were delivered in a traditional manner, we excluded the first MOOC from our study. We also excluded the third MOOC, given that the number of enrolled students was barely small. To sum up, our study in this paper considers only the second MOOC that covers the second part (weeks 5 to 8) of the flipped course. To pass the course, it is mostly necessary to obtain at least 60% of the total points for each assignment. Hence, we used this rule as a way to measure success in our study. The final data set consists of clickstream data from 170 students, with 33% of them failing the course.

Table 2: Features selected as important by RF classifiers for the ensemble of features for each course.

Set	Name	Short Description	LA-Flip	FP-Flip	LA-MOOC	FP-MOOC
AkpinarEtAl	QCheck-QCheck-QCheck	The amount of times the student checks three times the same quiz.		✓		
	QCheck-QCheck-VLoad	The amount of times the student checks twice the quiz and then go to load a video.	✓			
	VPlay-VPlay-VPlay	The amount of times the student clicks for three consecutive times on play for three different videos.			✓	
	VPlay-VPause-VLoad	The amount of times the student plays a video, pause and then load another one.	✓			
	VPlay-QCheck-QCheck	The amount of times the student plays a video, then checks twice a quiz.				✓
	VPlay-VStop-VPlay	The amount of times the student plays a video, stops, and plays another one.		✓		
	VPause-VSpeedChange-VPlay	The amount of times the student pauses a video, changes the speed, and re-plays it.		✓		
	VStop-VPlay-VSeek	The amount of times the student stops a video, then re-play it and seek to a given part.		✓		
	VStop-VCheck-VLoad	The amount of times the student stops a video, checks a quiz and then load another video.	✓			
BoroujeniEtAl	DelayLecture	The average delay in viewing video lectures, as soon as they are released.		✓	✓	✓
	RegWeeklySim-M1	The extent to which the student works on the same weekdays across weeks.	✓	✓		
	RegWeeklySim-M2	The extent to which a student has a similar distribution of workload among weekdays across weeks.	✓		✓	✓
	RegWeeklySim-M3	The extent to which the time spent on each day of the week is similar for different weeks of the course.		✓		
	RegPeriodicity-M1	The extent to which the hourly pattern of student's activities is repeating over days.				✓
	RegPeriodicity-M3	If the daily study pattern is repeating over weeks (e.g. is active on Monday and Tuesday in every week).	✓			✓
	RegPeakTime-M1	The extent to which students' activities are centered around a particular hour of the day.				✓
	RegPeakTime-M2	The extent to which students' activities are centered around a particular day of the week.				✓
ChenCui	RatioClicksWeekendWeekdays	The ratio between clicks in weekdays and weekends.	✓	✓	✓	✓
	TimeSession-Avg	The average amount of time spent from a login to the end of the session.			✓	
	TimeSession-Std	The standard deviation of time spent from a login to the end of the session.				✓
	TimeBetweenSessions	The average amount of time passed between two sessions for a student.				✓
	TotalClicks-Weekdays	The number of clicks performed by a student over weekdays.			✓	
LalleConati	PauseDuration-Avg	The average amount of time spent in pause while interacting with a video.		✓		
	SeekLength-Std	The extent to which the seek length varies across videos.		✓		
	PauseDuration-Std	The extent to which the pause duration varies across videos.		✓		
	TimeSpeedingUp-Avg	The average amount of time spent with higher than 1x speed while playing a video.		✓		
	TimeSpeedingUp-Std	The extent to which the time spent speeding up higher than 1x the videos varies.		✓		
	WeeklyPropWatched-Avg	The proportion of videos the student watched, counting repeating segments.	✓			
	WeeklyPropInter-Avg	The average time the student spent in interrupting a video.	✓			
	WeeklyPropInter-Std	The deviation of the time the student spent in interrupting a video.	✓			
	WeeklyPropReplayed-Avg	The proportion of videos the student re-watched, counting repeating segments.	✓			
	WeeklyPropReplayed-Std	The deviation of the proportion of videos the student re-watched, counting repeating segments.			✓	
	FrequencyEvent-VideoPlay	The frequency of the play event in the students' sessions.	✓			✓
MubarakEtAl	SpeedPlayBack-mean	The average speed the student used to play back a video.			✓	✓
WanEtAl	NumSubmissionsCor	The number of quizzes attempted and correct.	✓			
	NumSubmissions-Avg	The number of submissions performed for a quiz, on average.	✓			✓
	NumSubmissionPerCorrect	The percentage of the correct quiz submissions with respect to the total submissions.	✓		✓	
	NumSubmissionDistinct	The total number of distinct problems attempted by the student.			✓	
Ours	CompetencyAnticipation	The extent to which the student approaches soon a quiz provided in subsequent weeks.	✓			
	ContentAnticipation	The extent to which the student approaches soon a video provided in subsequent weeks.	✓			
	CompetencyStrength	The extent to which a student passes a quiz getting the maximum grade with a low number of trials.	✓		✓	
	StudentShape	The extent to which the student receives the maximum grade in quizzes at the first attempt in a row.	✓		✓	
	Student Speed	The average amount of time passed between two submissions for the attempted quizzes.			✓	

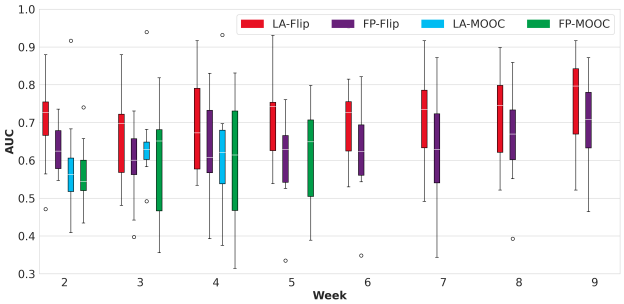


Figure 5: AUC scores per week for RF classifiers trained on feature ensembles. Flipped courses (*-Flip) last 10 weeks; LA-MOOC (FP-MOOC) last 4 (6) weeks.

Data Set: FP-MOOC. The content delivered in pre-class activities in FP-Flip was also provided by EPFL instructors on an external MOOC platform in form of two separate MOOCs, with the first MOOC being equivalent to the first 6 weeks of the flipped course and the second MOOC to the subsequent weeks. No data was available on the second MOOC, so we limited our study to only the first MOOC (week 1 to 6 of the flipped period of FP-Flip). To pass this MOOC, 80% of the total points for each of the five graded assignments are mostly needed. Hence, we used this rule to measure success in our study. The dataset consists of clickstreams from 3,565 students, with 52% failing the course.

5.2 Observations

We evaluated the predictive performance of a RF classifier across weeks for each course for the best ensemble feature set for that course. Figure 5 illustrates the predictive performance across weeks for all four courses. Considering the same course across different settings (flipped or MOOC), it can be observed that RFs trained on flipped course data achieved higher AUC scores than their MOOC counterpart. This difference can be due to multiple reasons, for example the different educational setting or the way the passing rule for the course is set up. Considering courses in the same setting (LA-Flip VS FP-Flip or LA-MOOC VS FP-MOOC), the results show that including quizzes in the LA-Flip course allows to increase the predictive power of the considered classifiers, compared to FP-Flip, that has no quizzes. This can be associated to the fact that quizzes are a good source of information for grasping the students' performance. The same observation is, however, less strong on the MOOC counterpart of the same two courses, highlighting again the high dependency from the educational setting.

In a second part of this experiment, we analyzed the average importance across weeks of the features selected by RFs across courses. Table 2 shows for each feature set and course, whether a given feature has been selected by the corresponding RF classifier. It should be noted that this table includes only features picked at least by a RF classifier across courses. In general, we show that while there is some overlap between the optimal features across courses, the importance of the features highly depends on the setting and structure of the course. The ratio of clicks between weekends and weekdays (ChenCui - RatioClicksWeekendWeekdays) is selected by all classifiers in all settings. Other features with a good level of generalizability are represented by those measuring regularity (BoroujeniEtAl). The other features were picked according to the setting or the structure of the course. In particular, RFs trained on LA-Flip and LA-MOOC assigned a higher importance to features that measure behavior in quizzes (e.g., Ours or WanEtAl). Hence, we can conclude that when available, features on quizzes are frequently selected, regardless of the setting. For courses with no quizzes, namely FP-Flip and FP-MOOC, the predictive power of RFs

is mainly based on regularity and fine-grained video usage (e.g., features on time spent in a video, e.g., LalleConati).

Highlight #5. *When quizzes are included in the schedule, quiz-related features are frequently selected as important. This is stronger in flipped than MOOC settings. When only videos are available, the predictive power mainly derives from regularity and fine-grained video-related features.*

For the same course in different settings, namely LA-Flip VS LA-MOOC and FP-Flip VS FP-MOOC, the optimal feature set heavily changed. In LA courses, quiz-related features were more important in the flipped context, while session-based features were more important in MOOCs (e.g., those from ChenCui). The latter finding holds for FP courses as well. Specifically, RFs trained on the MOOC version consistently selected features related to the students' session. Another observation for FP is that in the flipped version, tri-grams (ApkinarEtAl) and fine-grained video usage features (LalleConati) were picked; in the MOOC, regularity and session-based features were more important. To sum up, according to Table 2:

Highlight #6. *Predictors in flipped settings often rely on features based on tri-grams and fine-grained video consumption. Conversely, predictors in MOOCs consider regularity and session-based features as important. Quiz-related feature are picked in both settings, when quizzes are available.*

6. DISCUSSION

In this section, we connect the main findings coming from the individual experiments and present the implications and limitations of our study in the early success prediction task.

Course-Related Observations. A challenge, as our work shows, lies on the generalizability of feature predictive power across courses. The variability of the results when repeating the exact same experiment with data from different courses (or slightly different settings) is very high. It is therefore challenging to understand when, why, and how a feature tested on a given course could be re-used for other courses.

Highlight #7. *The predictive power of features does not often generalize across courses with different structures and educational settings. This observation is stronger with respect to the courses structure than between flipped and MOOC settings.*

This observation affects the scalability of early predictors. Being so course-dependent, identifying and enabling features predictive of student success for a given course can take hours or days, given that the intellectual and experimental work needs to be replicated on courses, case by case.

Highlight #8. *The lack of feature predictive power generalizability questions the extent to which a feature can be scaled across courses with the same structure/setting.*

Our experiments also showed that including quizzes in pre-class activities leads to substantial improvements in effectiveness. Hence, success prediction is driven by complex relationships between students' characteristics and the course domain, structure, and educational setting.

Data-Related Observations. Research in the area of early success prediction is often conducted on data extracted from online activities only. Even in our case study (for LA-flip), we could not rely on data collected in class, missing an important segment of learning. Moreover, clickstreams in this study do not cover other relevant interactions such as those in forums. In flipped courses, most (non-digitalized) discussions happen in class, and the forum is mainly used by teachers for announcements.

Highlight #9. *Early success prediction in flipped courses would benefit from including data coming from offline activities (e.g., in class).*

Workflow-Related Observations. To establish reproducibility, the description of the proposed features should go beyond plain-text only. Our formulation in this paper can be re-used to define features as formulas, making it easier to replicate them, especially when no source code is provided.

Highlight #10. *Feature descriptions can be accompanied by their mathematical formulation to ease reproducibility. When possible, sharing the code can facilitate their re-use.*

Though we validated the current features on RFs, other classifiers were not presented. However, RFs often provide the best trade-off between effectiveness and interpretability (the latter was fundamental for our study) and our framework makes it easy to run this analysis on other classifiers. Given that other classifiers (e.g., Support Vector Machines) gave worse (or comparable) results in the preliminary experiments we ran, our results depict a valid picture of feature predictive power.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed recent features for early success prediction in flipped and online courses. First, we investigated the predictive power of eight existing feature sets and a novel feature set proposed in this paper on a flipped course. We benchmarked the predictive power of features using a RF classifier, and discussed the ensemble feature set optimal for that course. We then extended our analysis to courses with other settings (MOOCs), domains, and structures, showing that the optimal ensemble and its predictive power vary. Our work calls for generalizable early predictors across courses with different characteristics. To promote research in this field, we also publicly release the source code developed during our study (see the footnote in Section 1).

In future work, we plan to extend our analysis to other features (e.g., based on in-class data), and types of student success tasks (e.g., grade prediction). We also plan to analyze more advanced classifiers and to devise robust classifiers across courses before testing them in the real world.

Acknowledgments We thank Himanshu Verma (TU Delft, formerly at EPFL) and Patrick Jermann and Francisco Pinto (EPFL Center for Digital Education) for the valuable input and support on the data sharing and manipulation.

8. REFERENCES

[1] N. Akpinar, A. Ramdas, and U. Acar. Analyzing student strategies in blended courses using clickstream data. In *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*, pages 6–17. Int. Educat. Data Mining Society, 2020.

[2] B. J. Beatty, Z. Merchant, and M. Albert. Analysis of student use of video in a flipped classroom. *TechTrends*, 63(4):376–385, 2019.

[3] M. S. Boroujeni, K. Sharma, L. Kidzinski, L. Lucignano, and P. Dillenbourg. How to quantify student's regularity? In *Proceedings of the 11th European Conference on Technology Enhanced Learning, EC-TEL 2016*, volume 9891 of *Lecture Notes in Computer Science*, pages 277–291. Springer, 2016.

[4] F. Chen and Y. Cui. Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, 7(2):1–17, 2020.

[5] J. Gardner and C. Brooks. Student success prediction in moocs. *User Modeling and User-Adapted Interaction*, 28(2):127–203, 2018.

[6] N. Goedhart, N. Blignaut-van Westrhenen, C. Moser, and M. Zweekhorst. The flipped classroom: supporting a diverse group of students in their learning. *Learning Environments Research*, 22(2):297–310, 2019.

[7] H. He, Q. Zheng, B. Dong, and H. Yu. Measuring student's utilization of video resources and its effect on academic performance. In *Proceedings of the 18th IEEE International Conference on Advanced Learning Technologies, ICALT 2018*, pages 196–198. IEEE Computer Society, 2018.

[8] J. Jovanović, D. Gašević, S. Dawson, A. Pardo, N. Mirriahi, et al. Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, 33(4):74–85, 2017.

[9] J. Jovanovic, N. Mirriahi, D. Gasevic, S. Dawson, and A. Pardo. Predictive power of regularity of pre-class activities in a flipped classroom. *Computers & Education*, 134:156–168, 2019.

[10] C. Lai and G. Hwang. A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Computers & Education*, 100:126–140, 2016.

[11] S. Lallé and C. Conati. A data-driven student model to provide adaptive support during video watching across moocs. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education, AIED 2020*, volume 12163 of *Lecture Notes in Computer Science*, pages 282–295. Springer, 2020.

[12] J. Lee and H. Choi. Rethinking the flipped learning pre-class: Its influence on the success of flipped learning and related factors. *British Journal of Educational Technology*, 50(2):934–945, 2019.

[13] D. J. Lemay and T. Doleck. Grade prediction of weekly assignments in MOOCS: mining video-viewing behavior. *Education and Information Technologies*, 25(2):1333–1342, 2020.

[14] G. S. Mason, T. R. Shuman, and K. E. Cook. Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course. *IEEE Transactions on Education*, 56(4):430–435, 2013.

[15] B. Mbouzao, M. C. Desmarais, and I. Shrier. Early prediction of success in MOOC from video interaction features. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education, AIED 2020*, volume 12164 of *Lecture Notes in Computer Science*, pages 191–196. Springer, 2020.

[16] A. A. Mubarak, H. Cao, and S. A. M. Ahmed. Predictive learning analytics using deep learning model in moocs' courses videos. *Education and Information Technologies*, 26(1):371–392, 2021.

[17] E. M. W. Ng. Integrating self-regulation principles with flipped classroom pedagogy for first year university students. *Computers & Education*, 126:65–74, 2018.

[18] J. O'Flaherty and C. Phillips. The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, 25:85–95, 2015.

[19] S. Park and N. H. Kim. University students' self-regulation, engagement and performance in flipped learning. *European Journal of Training and Development*, 2021.

[20] W. W. Porter, C. R. Graham, K. A. Spring, and K. R. Welch. Blended learning in higher education: Institutional adoption and implementation. *Computers & Education*, 75:185–195, 2014.

[21] A. A. Rahman, B. Aris, M. S. Rosli, H. Mohamed, Z. Abdullah, and N. Mohd Zaid. Significance of preparedness in flipped classroom. *Advanced Science Letters*, 21(10):3388–3390, 2015.

[22] M. Shih, J. Liang, and C. Tsai. Exploring the role of university students' online self-regulated learning in the flipped classroom: a structural equation model. *Interact. Learn. Environ.*, 27(8):1192–1206, 2019.

[23] N. A. Uzir, D. Gasevic, W. Matcha, J. Jovanovic, and A. Pardo. Analytics of time management strategies in a flipped classroom. *Journal of Computer Assisted Learning*, 36(1):70–88, 2020.

[24] J. N. Walsh and A. Risquez. Using cluster analysis to explore the engagement with a flipped classroom of native and non-native english-speaking management students. *The International Journal of Management Education*, 18(2):100381, 2020.

[25] H. Wan, J. Ding, X. Gao, and K. Liu. Supporting quality teaching using educational data mining based on openedx platform. In *Proceedings of the IEEE Frontiers in Education Conference, FIE 2017*, pages 1–7. IEEE Computer Society, 2017.

[26] H. Wan, K. Liu, Q. Yu, and X. Gao. Pedagogical intervention practices: Improving learning engagement based on early prediction. *IEEE Transactions on Learning Technologies*, 12(2):278–289, 2019.

[27] K. Wang and C. Zhu. Mooc-based flipped learning in higher education: students' participation, experience and learning performance. *International Journal of Educational Technology in Higher Education*, 16(1):1–18, 2019.

[28] R. M. Yilmaz and O. Baydas. An examination of undergraduates' metacognitive strategies in pre-class asynchronous activity in a flipped classroom. *Educational Technology Research and Development*, 65(6):1547–1567, 2017.